Is AI Instruction Comparable to Human Instruction? Designing a Pedagogical Agent for Complex Task Training

Bradford Schroeder, Jason Hochreiter, Sean Thayer,
Javier Rivera and Wendi Van Buskirk

January 18, 2025

Is AI Instruction Comparable to Human Instruction? Designing a Pedagogical Agent for Complex Task Training

Bradford L. Schroeder, Jason E. Hochreiter, Sean C. Thayer, Javier A. Rivera, and Wendi L. Van Buskirk

Naval Air Warfare Center Training Systems Division, Orlando, FL 32826, USA
{bradford.l.schroeder.civ,jason.e.hochreiter.civ,sean.c.thayer.civ,javier.a.rivera59.civ,wendi.l.vanbuskirk.civ}@us.navy.mil

Abstract. Complex tasks can be difficult to train and difficult to learn. Though training one-on-one with a human instructor is considered the gold standard for learning, social agency theory posits that social cues in virtual instructors can elicit the same mental processes in learners that occur with human instructors. We contend that artificial intelligence (AI)-driven pedagogical agents (PAs) could be superior for analyzing learner data, adapting instructional content, and providing scalable, consistent, on-demand training, particularly when instructors are expensive or unavailable. Ideally, an AI-driven PA would surpass the capability of a human instructor if these features could be fully realized. Because there are key differences in the way that humans and AI present learning content, we must ascertain whether a PA can deliver a quality of learning similar to a human instructor. We compared the knowledge levels of two groups after learning a complex task with either a human instructor or PA. We present nonexperimental preliminary findings that suggest learning from a human instructor and PA are comparable and discuss instructional considerations for PA design approach.

Keywords: Adaptive Training, Artificial Intelligence, Complex Learning, Human Performance, Virtual Instructors

1 Introduction

Human one-on-one instruction has long been the gold standard for delivering a flexible style of instruction that is not possible in the group setting [1]. While individualized instruction is considered optimal, it imposes a costly burden to training programs. Modern artificial intelligence (AI) solutions, such as pedagogical agents (PAs), could reduce costs without significant sacrifices in training quality. When PAs are supported by well-designed instructional algorithms, they can provide a tailored learning process similar to the one-on-one learning experience.

Social agency theory (SAT) [2] offers a practical theoretical framework to guide instructional algorithm design considerations. SAT posits that the use of social cues in a non-human entity (such as a PA) can trigger the same social conversational schema elicited during human instruction. For example, the integration of facial expressions, politeness, natural voices, and titular names might prime a human-nonhuman interaction to be perceived as inherently social [2, 3]. Training outcomes benefit from social schema activation because the learner engages a host of cognitive processes (e.g., cooperation principle, deep cognitive learning) to make sense of what the PA is communicating [4]. Perhaps unwittingly, the learner abides the social rules of the exchange while making lasting, meaningful representations of the information in memory [2, 5].

Nonetheless, designing effective instructional algorithms for training complex tasks can be challenging. For instance, if the task to-be-trained is complex in nature, it likely includes difficult decision-making and problem-solving aspects where learners often make mistakes. These mistakes are valuable for human instructors who use them to provide performance-based feedback and modify instruction to adapt to their learner's needs. Instructional algorithms may also be able to detect these errors and the reasons they were made. Using this information, PAs could deliver instruction through multimedia (e.g., text, images, audio clips, graphics, videos) and a variety of interventions (e.g., real-time feedback, after-action review, scaffolding) to approximate human instruction. However, determining when to present instructional materials, prioritizing competing learning goals, and ensuring the learner understands the PA are complicated aspects of designing effective instructional algorithms. Therefore, thoughtful design of a comprehensive PA algorithm must account for a variety of learner needs and performance outcomes in a complementary way.

To this end, we created a dynamic instructional algorithm to compare a PA to the gold standard of one-on-one instruction for learning outcomes in a complex task. Previous work has explored PA algorithms as part of computer-based instruction and adaptive training (AT) delivering end-to-end complex task instruction [6, 7, 8]. This research suggests that well-designed PA algorithms support maintenance of learning and improve on-task performance [9]. However, it remains to be seen whether PA algorithms can meet (or even exceed) the learning benefits of human instruction.

2 Methodology

We performed two data collections where we trained groups of participants on a complex radio frequency (RF) identification task. The human instructor group (HI; data collection 1; $N = 90$) received one-on-one training from an experimenter and the PA group (PA; data collection 2; $N = 89$) received training from a PA instructor. During the task, participants monitored an environment for RF signals, performed signal analysis, and prepared and submitted time-sensitive reports. Both groups received training on how to perform the RF identification task, including using the interface (taking measurements, finding information, and submitting reports) and signal analysis (understanding parameter information, priorities of each signal type, reporting requirements, etc.). After the training session, both groups answered 10 multiple-choice quiz questions relating to these task topics.

To mimic the capabilities of a human instructor, we designed a PA that automatically monitored a participant's actions and provided them timely, relevant instruction through visual cues and audio clips via a feedback algorithm. For example, where a human instructor would have directed the learner's attention to relevant information via gesture, verbal direction, or controlling the learner's cursor, the PA would highlight the information, play a sound clip, or take control of the learner’s cursor. The design of the PA was informed in 2 ways: (1) using SAT as a framework and (2) by our own protocol for providing human instructor-led training for this task. Regarding SAT, we designed a PA to present as an animated instructor, Captain Ray, with a human-like face and titular name. Through human voice recordings, Captain Ray provided feedback in a polite, conversational voice (vs. machine synthesized) to sustain social credibility throughout the training session. We also instantiated a gesture analog by using on-screen visual cues and cursor movements to simulate human-like pointing and other guiding gestures. In essence, the PA served as an “over-the-shoulder” instructor, governed by algorithms that emulate the instructional sequences a human instructor would provide to the learner. Regarding the human instructor protocol, there were two types of feedback delivered to learners: assistive and corrective. Assistive feedback guided the learner when they needed their attention redirected (e.g., “Check the new signal that came in”), prompting them to attend to the next part of the task they needed to complete. Corrective feedback addressed an error on a numerical or categorical response (e.g., “The measurement you just provided was incorrect; remember to measure the full waveform”). In both groups, assistive or corrective feedback was provided as needed, determined by either the experimenter’s judgment of learner performance (HI group) or driven by the PA algorithm (PA group). The content of the PA group’s feedback algorithm was generated based on scripts for training learners in the HI group.

For the PA algorithm, we designed a set of prioritized feedback rules, each with three major components: the conditions under which feedback should be presented, the specific content to be presented (e.g., imagery or audio), and the conditions under which the feedback should be removed (e.g., the learner addressed their mistake). Each of these rules provided either assistive or corrective feedback to the learner based on their actions, inaction, or at scheduled times during training. Over time, the feedback algorithm monitored the learner's performance, queuing up any rules for which the feedback presentation conditions were satisfied and displaying them in priority order. As an example, participants were responsible for monitoring the environment for the appearance of new RF signals. Given the inherent complexity of the task and the interface, these signal appearance events were often easy to overlook. To address this, we designed a feedback rule to detect whether a new signal had appeared that the learner had not attended to after a few seconds had elapsed. If so, the algorithm played audio from the captain informing the learner that they had missed the new signal and provided visual cues guiding their attention toward it. Importantly, the scripts, imagery, and timing of the algorithm were fixed based on performance for every participant in the PA group. This is a contrast to the HI group, where instructors provided verbal feedback based on learner mistakes following the same script, but the timing varied naturally with each session.

3 Results

We compared both groups’ scores on a knowledge test following their tutorial session using the two one-sided t-test (TOST) procedure described by Lakens et al. [10]. The result was significant, $t(177) = -1.76$, $p = .04$, providing evidence that both versions of instruction had statistically equivalent overall quiz scores (M_{HI} = 90.44%, SD_{HI} = 10.70%; M_{PA} = 87.75%, SD_{PA} = 12.95%). Although this result supported our goal, we wanted to explore frequency data with feedback delivered by the algorithm and how it may have related to knowledge test scores in the PA group. Though those in the HI group received the same types of feedback, the algorithm automatically collected data relating to when and why specific feedback was shown, which could offer insight for improving the algorithm and understanding how it affected learning. Eighty-five participants' algorithm data were available for analysis.

First, we assessed the frequency that each of our 24 rules was triggered by participants. Examining these data offers a perspective for how often the typical participant triggered feedback from the PA, which might suggest areas for algorithm improvement (for example, rules that were rarely triggered by participants may need to be revised or discarded). Among this sample, all rules were triggered at least once, and some were triggered more than once for the average participant (see Figure 1). Next, we examined differences between assistive and corrective feedback to understand whether the PA was providing more of one type of feedback or another and how this may relate to knowledge acquisition. There were a total of 9 corrective feedback rules and 15 assistive feedback rules. Between these two types, 22% of feedback delivered was corrective, and 78% was assistive. Interestingly, the total sum of feedback triggers was correlated with performance on the knowledge test, $r(83) = -.28$, $p = .009$, such that more feedback interventions from the PA resulted in lower scores on the knowledge test. This was primarily driven by corrective feedback, $r(83) = -.41$, $p < .0001$, such that those who received more corrective feedback performed more poorly on the knowledge test. The correlation for assistive feedback was not significant, $r(83) = -.12$, $p = .27$.

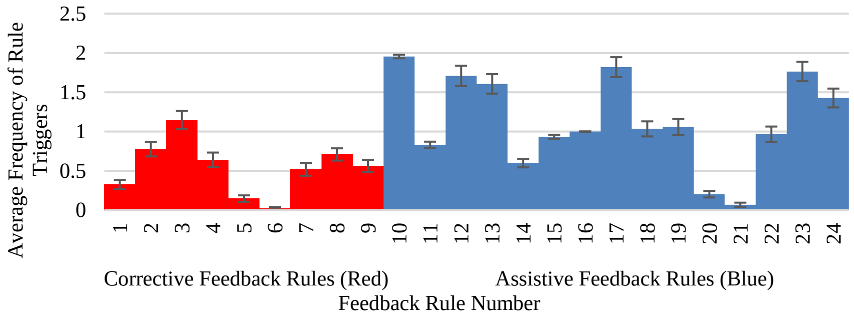

Figure 1. Frequency diagram of feedback algorithm rule triggers. Error bars = SE.

4 Discussion

Our observed result was encouraging, but we contend that there are still areas for improvement, and our results have limitations. First, we acknowledge that multiple-choice quiz scores only provide a partial account of declarative knowledge acquired from instruction. Assessing procedural knowledge and real-time decision-making performance may elucidate more differences from these two modes of instruction.

In evaluating our algorithm, we observed that every feedback rule was triggered; however, this could indicate that we neglected to design other feedback-worthy rules. In addition, the rules we did employ may have had some limitations for learning. For example, when the PA provided feedback, it assumed the participant understood it when they addressed the underlying mistake that prompted it; participants were unable to seek clarification. Anecdotally, some participants ignored or missed the PA's audio-visual feedback, perhaps assigning it less importance than if received from a human instructor. The high workload from the task itself may have made it difficult for participants to attend to this feedback, especially if it interrupted their workflow or was not as salient as feedback from a human instructor. Instead, a human instructor might be better able to deliver feedback at more opportune times. It may also be possible to increase the saliency of visual feedback to ensure that it is not missed. For example, the PA could guide a learner's attention by tracking their cursor or eye movements and creating a visual path toward relevant feedback. Though we endeavored to account for a variety of possible learner interactions, we were unable to accommodate them all using this prescribed feedback, an ability human instructors can often perform well.

In general, the PA instructor provided more assistive feedback than corrective feedback, but corrective feedback was more strongly and negatively correlated with performance on the knowledge test. Future research is needed to understand why this happened. Were participants who received more corrective feedback unable to properly encode the task procedure because corrective feedback interrupted their flow to correct their work? Or did they poorly encode the instructional material prior to training with the PA? Given that corrective feedback sought to address specific signal classification errors, it may be the case that a participant who received less of this feedback type understood the task sufficiently well and therefore performed well on the knowledge test. Additionally, this poses the question of how this corrective feedback could be improved to minimize the performance gap between those who committed more mistakes and required more of this type of feedback and those who achieved high scores without it. Seeing equivalence between a human instructor and PAs designed utilizing SAT principles is promising. However, we do not yet believe it fully capitalizes on the benefits of AI-driven PAs, such as scalability and performance data analysis. With these shortcomings addressed, PAs could exceed the training capabilities of a one-on-one human instructor.

5 Conclusion

This work describes an exploratory analysis comparing the effectiveness of an AI instructor against a human instructor for training a complex RF identification task and its effects on learning. Results suggested both approaches provided instruction to learners equivalently; however, future research is necessary to determine the efficacy of using PAs and associated feedback algorithms in other task domains and with other measures of learning to better bridge the gap between them and human instructors.

Acknowledgments. This work was funded by the Naval Innovation Science and Engineering program. Presentation of this material does not constitute or imply its endorsement, recommendation, or favoring by the U.S. Navy or the Department of Defense (DoD). The opinions of the authors expressed herein do not necessarily state or reflect those of the U.S. Navy or DoD. NAWCTSD Public Release 24-ORL040 Distribution Statement A – Approved for public release; distribution is unlimited.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bloom, B. S.: The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational researcher 13 (6), 4-16 (1984).
2. Mayer, R. E., Sobko, K., Mautone, P. D.: Social cues in multimedia learning: Role of speaker's voice. Journal of Educational Psychology, 95 (2), 419 (2003).
3. Reeves, B., Nass, C.: The media equation: How people treat computers, television, and new media like real people. Cambridge, UK, 10 (10) (1996).
4. Grice, H. P.: Logic and conversation. In: Cole, P., Morgan, J. (eds.), Syntax and semantics, vol. 3, pp. 41–58. Academic Press, New York. (1975).
5. Mayer, R. E.: Multimedia learning. Cambridge University Press, New York (2001).
6. Landsberg, C. R., Mercado, A. D., Van Buskirk, W. L., Lineberry, M., Steinhauser, N.: Evaluation of an adaptive training system for submarine periscope operations. In: Proceedings of the Human Factors and Ergonomics Society annual meeting, vol. 56, No. 1, pp. 2422-2426. Sage Publications, Los Angeles, CA (2012).
7. Schroeder, B. L., Fraulini, N. W., Van Buskirk, W. L., Miller, R. M.: Assessing the Social Agency of Pedagogical Agents in Adaptive Training Systems. In: Sottilare, R.A., Schwarz, J. (eds.) Adaptive Instructional Systems, HCII 2022, LNCS, vol. 13332, pp. 302-313. Springer International Publishing, Switzerland (2022).
8. Schroeder, B. L., Van Buskirk, W. L., Aros, M., Hochreiter, J. E., & Fraulini, N. W.: Which is Better Individualized Training for a Novel, Complex Task? Learner Control vs. Feedback Algorithms. In: Sottilare, R.A., Schwarz, J. (eds.) Adaptive Instructional Systems, HCII 2023, LNCS, vol. 14044, pp. 236-252. Springer International Publishing, Switzerland (2023).
9. Schroeder, B. L., Fraulini, N. W., Van Buskirk, W. L., Johnson, C. I.: Using a Non-player Character to Improve Training Outcomes for Submarine Electronic Warfare Operators. In: Sottilare, R.A., Schwarz, J. (eds.) Adaptive Instructional Systems, HCII 2020, LNCS, vol. 12214, pp. 531-542. Springer International Publishing, Switzerland (2020).
10. Lakens, D., Scheel, A. M., Isager, P. M.: Equivalence Testing for Psychological Research: A Tutorial. Advances in Methods and Practices in Psychological Science 1 (2), 259-269 (2018).